AD-A077 296 MASSACHUSETTS INST OF TECH CAMBRIDGE LAB FOR INFORMA--ETC F/G 17/2
THE ESTIMATION OF DELAY GRADIENTS FOR PURPOSES OF ROUTING IN DA--ETC(U)
SEP 79 M G BELLO , A SEGAL N00014-75-C-1183
UNCLASSIFIED LIDS-P-939 NL

| OF |
ADA
077296

END
DATE
FILMED
12-79
DDC

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)

THE ESTIMATION OF DELAY GRADIENTS FOR PURPOSES OF ROUTING IN DATA-COMMUNICATION NETWORKS

5. TYPE OF REPORT & PERIOD COVERED

Paper

6. PERFORMING ORG. REPORT NUMBER

LIDS-P-939

7. AUTHOR(s)

M.G. Bello
Adrian Segal

8. CONTRACT OR GRANT NUMBER(s)

ARPA Grant N00014-75-C-1183

9. PERFORMING ORGANIZATION NAME AND ADDRESS

M.I.T.
Laboratory for Information and Decision Systems
Cambridge, MA 02139

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

Program Code No. 5T10
ONR Identifying No. 049-383

11. CONTROLLING OFFICE NAME AND ADDRESS

Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

12. REPORT DATE

September 1979

13. NUMBER OF PAGES

48

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

Office of Naval Research
Information Systems Program
Code 437
Arlington, Virginia 22217

15. SECURITY CLASS. (of this report)

Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

(15) N00014-75-C-1183

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

DDC RECEIVED NOV 27 1979 A

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The derivatives with respect to message flow of the total delay accumulated per unit time on each link in a Data-Communication network have been shown to be fundamental quantities in the solution of the routing problem. Casting the problem of estimating these delay gradients in a queueing theory framework and making no statistical assumptions other than stationarity, we propose three algorithms that process the record of arrivals and departures of a single-server queue to derive an estimate for the derivative, with respect to arrival rate, of the total delay accumulated per unit time. Through simulation and

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD A077296

DDC FILE COPY

79 11 26 096

410 950 SB

analysis we show that all three algorithms are asymptotically unbiased and efficient for M/D/1 queues. By simulation of other queues we investigate the relative robustness of the three procedures. Finally, though examination of the storage and computational requirements we identify a single most promising algorithm.

LIDS-P-939 September 1979

THE ESTIMATION OF DELAY GRADIENTS FOR PURPOSES OF ROUTING IN DATA-COMMUNICATION NETWORKS*

M.G. Bello**

and

Adrian Segal***

ABSTRACT

The derivatives with respect to message flow of the total delay accumulated per unit time on each link in a Data-Communication network have been shown to be fundamental quantities in the solution of the routing problem. Casting the problem of estimating these delay gradients in a queueing theory framework, and making no statistical assumptions other than stationarity, we propose three algorithms that process the record of arrivals and departures of a single-server queue to derive an estimate for the derivative, with respect to arrival rate, of the total delay accumulated per unit time. Through simulation and analysis we show that all three algorithms are asymptotically unbiased and efficient for M/D/1 queues. By simulation of other queues we investigate the relative robustness of the three procedures. Finally, through examination of the storage and computational requirements we identify a single most promising algorithm.

* This research was conducted in the MIT Laboratory for Information and Decision Systems with partial support provided by ARPA under grant #ONR-N00014-75-C-1183.

**Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

***Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 and Institute of Technology, TECHNION, Haifa, Israel.

1. INTRODUCTION

In several recent works [1]-[6] it has been shown that routing in a data-communication network, where the goal is to improve and minimize network average delay, should be based on knowledge of the derivative of delay on the network links with respect to the link flows. To be precise, if we let λ_{ik} be the average message flow (in messages/sec.) on a link (i,k) of a communication network and assume that the average delay faced by all messages going through this link is a function only of λ_{ik}, then we can write $D_{ik}(\lambda_{ik})$ for the total delay/unit time on the link, namely the product of the average delay per message and the average flow of messages/unit time. With these notations, several of the centralized and distributed routing algorithms introduced in the above mentioned references employ the derivative $D'_{ik}(\lambda_{ik}) = dD_{ik}/d\lambda_{ik}$ as a basic quantity. One way to obtain this quantity is to make some statistical assumptions on the message interarrivals and lengths and use one of the queueing formulae to obtain an explicit relationships between D_{ik} and λ_{ik}. For example, for an M/M/1 queue one obtains $D_{ik}(\lambda_{ik}) = \lambda_{ik}/(\mu_{ik}C_{ik} - \lambda_{ik})$ and now by measuring the flow λ_{ik}, the bit capacity C_{ik}, and $1/\mu_{ik}$, the average message length in bits, one can find $D'_{ik}(\lambda_{ik})$. Of course, different statistical assumptions will give different answers for $D'_{ik}(\lambda_{ik})$, and in fact these explicit formulae can be obtained only for the situation where independence and identical distribution of the variables involved is assumed. To circumvent some of these problems a different approach can be taken. Rather than estimate the flow and substitute into some

explicit formula for $D'_{ik}(\lambda_{ik})$, we can instead find algorithms to directly estimate $D'_{ik}(\lambda_{ik})$, the latter being independent of any statistical assumptions. The purpose of this paper is to introduce and analyze three procedures for estimating the derivative $D'(\lambda)$ of the total delay $D(\lambda)$ on a link as a function of the message flow rate λ on this link. The algorithms are based on measurements of arrival-departure records of the messages and on processing these quantities recursively to obtain the required estimates. We use the nomenclature of queueing theory by referring to the link as a queue and to the messages as customers.

In Section 2 we introduce and motivate the three procedures and present their realization in recursive form. Section 3 sketches the analysis of the statistical properties of each algorithm as applied to M/D/1 and M/G/1 queues. In Section 4 we compare and contrast the computational and storage requirements of each procedure. Finally, Section 5 presents the results of simulating the algorithms for M/D/1, M/M/1, U/M/1, and D/M/1 queues, and identifies a single most promising procedure.

2. THREE ESTIMATION ALGORITHMS

We now proceed to propose and evaluate three algorithms which process the record of a single server queueing system to estimate the derivative of the total delay/unit time with respect to arrival rate λ. The available record consists of exact knowlecge of arrival times of customers to the queue and of their departure times after service is completed. Time is segmented into alternative (random) intervals, busy periods, during which

the server is occupied, and idle periods, when the server is free. The observation interval which is used to form our estimate consists of a number of busy periods and the intervening idle periods.

A simple thought-experiment motivates all three estimation algorithms. Consider our single-server queueing system with its average arrival rate of λ customers per unit time. For a given observation period T_E, if we can compute the total system time S, i.e., the sum of all customer's service and waiting times, then the average delay/unit time is given by $D = S/T_E$. Suppose we could actually alter the input flow by some $\delta\lambda$. Then, on the basis of an earlier D, by computing D* for the next observation period, we can estimate the derivative of the total delay/unit time by calculating

$$\hat{D}' = \frac{D^*-D}{\delta\lambda} \quad . \tag{2.1}$$

However, in any actual queueing system it would be undesirable to change flows just for measurement purposes. Even if we could implement (2.1), the independent statistical fluctuations in D and D* would probably make it a very poor estimator. Hence, what we need is some mathematical formalism for an imaginary increment in flow $\delta\lambda$, which will allow us to compute the corresponding change in delay without actually perturbing the arrival rate.

According to intuition, an increase in arrival rate should result in additional customers entering the system. An extra customer arriving in a time interval T_E with probability ε will increase the effective rate by $\delta\lambda = \varepsilon/T_E$. If extra arrivals are mutually independent events, the

probability of two or more customers will be of second-order in ε and hence of second-order in $\delta\lambda$. Therefore, only the effect of a single extra arrival has to be considered explicitly. We also assume that the arrival time of the extra customer is uniformly distributed over the observation period T_E. In addition, in order to explicitly compute the change in total system time over the observation interval due to an extra arrival, we will assume that the additional customer has some known deterministic service requirement. These assumptions allow us to compute an expected increase in system time conditioned on the arrival of a new customer, and the resulting estimation procedure will be called the customer-addition algorithm.

In a second algorithm, first proposed in [1], an incremental decrease in the effective rate λ is simulated. This is done by assuming that each customer arriving to the system is allowed to indeed enter the queue with probability $1 - \varepsilon$, and is eradicated with probability ε, independently from customer to customer. In this way we simulate an arrival process with rate $\lambda(1 - \varepsilon)$, when $\varepsilon = -\delta\lambda/\lambda$. If we let M' be the total number of customers in a period T_E, then ε is given by

$$\varepsilon = \frac{-T_E}{M'} \delta\lambda \ . \tag{2.2}$$

Again, the probability of removal of two or more customers from the same period T_E is second-order in $\delta\lambda$ and hence the reduction of total system time that has to be considered explicitly is due to removal of

three algorithms will occur at the end of busy periods, and assumes the form

$$\hat{D}'(k) = \xi_k \hat{D}'(k-1) + \Delta_k. \tag{2.4}$$

Here, $\hat{D}'(k)$ denotes the delay derivative estimate based on k busy periods of observation. The factor ξ_k renormalizes $\hat{D}'_{(k-1)}$ to correspond to a term in the expression for the k-busy period estimate. Finally, Δ_k denotes the contribution to the k-busy period estimate derived from the current busy interval.

In the customer-addition algorithm we simulate an increase of $\delta\lambda$ in the arrival rate. The following assumptions will be made:

1) The probability of an extra arrival in the interval T_E is $\delta\lambda \cdot T_E$.
2) Each extra arrival is independent of all other arrivals.
3) The extra arrival is uniformly distributed over the interval T_E.
4) The service requirement of the extra customer is known; we denote it by x.

We must note at the start that the form of the customer-addition algorithm derived here is only meaningful for queues where all customers have identical service requirements. However, the procedure can be practically extended to the case where the service time distribution allows a finite set of discrete service requirements. For simplicity, we treat here the case of a single allowable service requirement.

Let T_k and I_k denote the busy and idle periods, respectively. Let $\delta s_{(N)}$ denote the increase in system time over N busy periods associated with the arrival of an extra customer. We let $\delta\hat{s}_{(N)}$ denote the expected increase in system time associated with an increase in arrival rate $\delta\lambda$ and conditioned on the record of arrivals and departures. By conditionining on the random arrival time t being in each T_k and I_k we can compute $\delta\hat{s}_{(N)}$ to first order in $\delta\lambda$ by

$$\delta\hat{s}_{(N)} = \{ \sum_{k=1}^{N} E(\delta s_{(N)} | t \in T_k, \text{ queueing record}) \frac{T_k}{T_E} + \sum_{k=1}^{N} E(\delta s_{(N)} | t \in I_k, \text{ queueing record}) \frac{I_k}{T_E} \} T_E \cdot \delta\lambda \qquad (2.5)$$

the factor $(T_E \cdot \delta\lambda)$ outside the brackets is the probability of an extra arrival. The factors T_k/T_E and I_k/T_E represent the probabilities of t being in the k-th busy and idle periods, respectively. This is a consequence of the assumption that t is uniformly distributed over T_E. Since we are interested in the derivative of the total delay/unit time with respect to flow rate, our estimator is given by

$$\hat{D}'_{(N)} = \frac{1}{T_E} \frac{\delta\hat{s}_{(N)}}{\delta\lambda} \quad . \qquad (2.6)$$

We focus next on the calculation of $E(\delta s_{(N)} | t \in T_k, \text{ queueing record})$ and $E(\delta s_{(N)} | t \in I_k, \text{ queueing record})$. These expected increments in system time are composed of the average effect of the extra arrival on existing customers plus the average system time of the additional customer. In considering additional arrivals in a busy period, we can distinguish

between the effects on customers in that busy period and interactions with succeeding busy periods. First we examine the part of $E(\delta s_{(N)} | t \in T_k,$ queueing record), call it Δs_k, that comes from considering the k-th busy period in isolation. To facillitate discussion, the following notation is defined:

$\tau_i \triangleq$ arrival time of i-th customer in busy period relative to start of busy interval.

$x_i \triangleq$ service requirement of the i-th customer

$s_i \triangleq$ system time of the i-th customer

$M \triangleq$ number of customers in the busy period

$T \triangleq$ duration of busy period.

The system time of a given customer is equal to his service plus waiting time. The waiting time is equal to the sum of the service requirements of those who enter the busy period before him minus his arrival time. Hence, for an arbitrary queue, the system time of the i-th customer is given by

$$s_i = x_i + \left(\sum_{\ell=1}^{i-1} x_\ell - \tau_i \right) \tag{2.7}$$

Now we note that an additional arrival at time t in the interval $[\tau_i, \tau_{i+1}]$ will have to wait $\left(\sum_{\ell=1}^{i} x_\ell - t \right)$ until being served and will cause the (M - i) following customers to wait an additional time x. Assuming that all the service requirements are the same ($x_\ell = x$), an arrival at time t results in an additional total waiting time of $(M - i)x + ix - t$ or

$Mx - t$. Now letting t be a uniformly distributed random variable on 0 to T or equivalently on 0 to Mx, we note that $\bar{t} = T/2$ or $\bar{t} = Mx/2$. Hence, the expected increment in system time Δs_k is given as

$$\Delta s_k = x + \Delta w_k \ , \tag{2.8}$$

where

$$\Delta w_k = Mx/2 \tag{2.9}$$

We complete the calculation of $E(\delta s_{(N)} | t \in T_k$, queueing record) by considering the additional system time that may result from one busy period overlapping onto another. No matter where an additional customer arrives in the k-th busy period, that period will be extended by the service time x. The value of x relative to the following idle period durations will determine the number of succeeding busy periods that will be affected by an arrival in T_k. If $x \leq I_k$ no following busy periods suffer additional delay. If $I_k \leq x \leq I_k + I_{k+1}$ only the (k+1)-st busy period is affected. The exact effect on a given busy period j in the future, depends on how much an arrival in T_k causes the (j-1)-st busy period to overlap into the j-th busy period. For example, if $x > I_k$, then each customer in T_{k+1} will suffer an additional delay $(x - I_k)$. Letting M_k denote the number of customers served in the k-th busy period, the preceding reasoning leads to the following rule for computing $E(\delta s_{(N)} | t \in T_k$, queueing record):

$$(\delta s_{(N)} \,|\, t \in T_k,\ \text{queueing record}) = \begin{cases} \Delta s_k & x \le I_k \\ \Delta s_k + M_{k+1}(x - I_k) & I_k \le x \le I_k + I_{k+1} \\ \vdots & \vdots \\ \Delta s_k + \sum_{j=1}^{\ell+1} M_{k+j}\left(x - \sum_{m=0}^{j-1} I_{k+m}\right) & \sum_{j=0}^{\ell} I_{k+j} \le x \le \sum_{j=0}^{\ell+1} I_{k+j} \\ \vdots & \vdots \\ \Delta s_k + \sum_{j=1}^{N-k} M_{k+j}\left(x - \sum_{m=0}^{j-1} I_{k+m}\right) & \sum_{j=0}^{N-k-1} I_{k+j} \le x \end{cases} \tag{2.10}$$

To complete our description of the customer-addition algorithm, we must now evaluate the average increase in system time $E(\delta s_{(N)} \,|\, t \in I_k,$ queueing record) associated with arrivals in idle periods. The effect of an arrival in I_k on the (j+k)-th busy period again depends on how much the (j+k-1)-st busy period slides onto the (j+k)-th busy interval, and thus must be averaged over all times of arrival t of the additional customer in the k-th idle period. Let t' denote the negative of the time between the arrival of the additional customer and the start of the (k+1)-st busy interval. Then, assuming an additional customer arrives in I_k, t' is a random variable uniformly distributed on the interval $[-I_k, 0]$. Let us define the quantity $\sigma_{k,j}$ by:

$$\sigma_{k,j} \triangleq \begin{cases} 0 & j=1 \\ \sum_{m=1}^{j-1} I_{k+m} & j \neq 1 \end{cases} \tag{2.11}$$

Then, an additional arrival in I_k will cause the (j+k-1)-th busy interval to overlap onto the (j+k)-th busy period by an amount equal to $\max\{x + t' - \sigma_{k,j}, 0\}$. Hence, each customer in the (j+k)-th busy period suffers an additional waiting time of $\max\{x + t' - \sigma_{k,j}, 0\}$. On the basis of this reasoning, we may compute $E\{\delta s_{(N)} \mid t \in I_k, \text{ queueing record}\}$ by

$$E\{\delta s_{(N)} \mid t\in I_k, \text{ queueing record}\} = x + \sum_{j=1}^{\ell+1} M_{k+j} \int_{t'=-I_k}^{0} \max\{x+t'-\sigma_{k,j},0\}\frac{dt'}{I_k} \tag{2.12}$$

for

$$\sum_{j=1}^{\ell} I_{k+j} \leq x \leq \sum_{j=1}^{\ell+1} I_{k+j} \tag{2.13}$$

The integrals appearing in relation (2.12) are computed as

$$M_{k+j}\int_{t'=-I_k}^{0} \max\{x+t'-\sigma_{k,j},0\}\frac{dt'}{I_k} = \begin{cases} M_{k+j}\left(x-\frac{I_k}{2} - \sigma_{k,j}\right) & (-I_k > \sigma_{k,j}-x) \\ \frac{M_{k+j}}{2I_k}\left(x-\sigma_{k,j}\right)^2 & (\sigma_{k,j}-x > -I_k) \end{cases} \tag{2.14}$$

Employing relations (2.8), (2.10), and (2.12) in (2.5) and (2.6) we can conceive of a processor which updates an estimate for the delay gradient at the end of each busy period. Let ℓ_k denote the time from the start of the observation period to the end of the k-th busy interval. Let Δ_k denote the incremental expected waiting time suffered by the M_k customers in the most recent busy period due to an additional arrival during the entire current queueing record plus the expected waiting time of an extra

customer arriving in T_k. Then, the estimate updating at the end of the k-th busy interval assumes the form

$$(\hat{D}'_{(k)}-x) = \frac{\ell_{k-1}}{\ell_k} (\hat{D}'_{(k-1)}-x) + \Delta_k \quad . \tag{2.15}$$

the $(\hat{D}'_{(i)}-x)$ terms appear since we can identify $(\hat{D}'_{(i)}-x)$ as an estimate of the incremental increase in waiting time per unit time.

Now, Δ_k can be expressed as

$$\Delta_k = (\frac{T_k}{\ell_k}) \Delta w_k + \Delta_{Ik} + \Delta_{Bk} \ , \tag{2.16}$$

where Δw_k denotes the expected additional waiting time suffered by members of the current busy period due to an additional arrival in T_k plus the newcomers expected waiting time, while Δ_{Ik} and Δ_{Bk} denote the expected waiting time suffered by customers in the most recent, k-th, busy interval due to an arrival in either past idle periods or busy periods respectively. Since x is finite, we need only look back a finite number of busy and idle periods to compute Δ_{Bk} and Δ_{Ik}. Let n_k denote the index of the first idle period at which a new arrival with service requirement x can affect the customers of the current, k-th, busy interval. We determine n_k by requiring that

$$\sum_{\ell=n_k+1}^{k-1} I_\ell < x < \sum_{\ell=n_k}^{k-1} I_\ell \quad . \tag{2.17}$$

Then, we may employ relations (2.10) and (2.12) to compute Δ_{Bk} and Δ_{Ik} by

$$\Delta_{Bk} = \sum_{\ell=n_k+1}^{k-1} M_k \left(x - \sum_{j=\ell}^{k-1} I_j\right) \left(\frac{T_\ell}{\ell_k}\right) \quad , \tag{2.18}$$

and

$$\Delta_{Ik} = \sum_{\ell=n_k}^{k-1} \left\{ M_k \int_{t'=-I_\ell}^{0} \max\{x+t'-\sigma_{\ell,k-\ell},\, 0\} \frac{dt'}{I_\ell} \right\} \frac{I_\ell}{\ell_k} \quad . \tag{2.19}$$

We note that at the end of each busy interval, n_k must be updated so that relation (2.17) is satisfied.

In the customer-removal algorithm we simulate a decrement in arrival rate $\delta\lambda$ by removing customers from the queue with probability ε (given by relation (2.2)) and computing the resulting decrement in total system time. Here, the situation is somewhat simpler than in the customer-arrival algorithm, since removal of a customer affects only the customers belonging to the same busy period. Letting $\delta s_{j,i}$ denote the change in the system time of the i-th busy period due to the removal of the j-th customer, the expected change in system time $\delta s_{(N)}$ in a N busy period observation interval T_E due to a decrement in flow $\delta\lambda$ and conditioned on the queueing record is formulated as

$$E(\delta s_{(N)} \,|\, \text{queueing record}) = \sum_{i=1}^{N} \sum_{j=1}^{M_i} - \delta s_{j,i} \left(\frac{T_E \delta\lambda}{M'}\right) \quad . \tag{2.20}$$

Hence, the desired delay gradient estimator is given by

$$\hat{D}'_{(N)} = \frac{1}{T_E} \frac{E(\delta s_{(N)} \,|\, \text{queueing record})}{\delta\lambda} = \frac{1}{M'} \sum_{i=1}^{N} \sum_{j=1}^{M_i} - \delta s_{j,i} \quad . \tag{2.21}$$

We compute $\delta s_{j,i}$ by working with more microscopic quantitites. Let $C^n_{m,i}$

denote the amount of system time saved for the m-th customer in the i-th busy period by the removal of the n-th customer in that busy period. Since the removal of the n-th customer can have no effect on customers that preceded him, $C^n_{m,i} = 0$ for m = 1,...,n-1. Hence, $\delta s_{j,i}$ can be computed as

$$\delta s_{j,i} = - \sum_{m=j}^{M_i} C^j_{m,i} \quad . \tag{2.22}$$

We now develop a systematic procedure for calculating the $C^n_{k,i}$'s. To simplify notation, we drop the i denoting the index of the busy period. Let w_n, s_n, and x_n denote the waiting time, system time and service requirement, respectively of the n-th customer in the busy period. Let d_n and a_n denote the corresponding departure and arrival time of the n-th customer. Since the system time the n-th customer saves by its own removal is s_n, we have $C^n_n = s_n$. In considering the effect of removing the n-th customer on the (n+1)-st customer, either a new busy period begins with the (n+1)-st or the (n+1)-st customer remains part of the busy period formed by customers 1 to n-1. The condition for customer n+1 beginning a new busy period is that the arrival time a_{n+1} of the (n+1)-st customer is greater than the departure time d_{n-1} of the (n-1)-st customer. In this case, customer n+1 will save its waiting time w_{n+1}. If $d_{n-1} > a_{n+1}$, removal of customer n does not cause customer n+1 to start a new busy period, and hence customer (n+1) saves an amount of time x_n since it need no longer wait for customer n to be served. This rule for C^n_{n+1} is summarized by

$$C^n_{n+1} = \begin{cases} d_n - a_{n+1} = w_{n+1} & \text{for } a_{n+1} \geq d_{n-1} \\ d_n - d_{n-1} = x_n & \text{for } d_{n-1} > a_{n+1} \end{cases} \tag{2.23}$$

Examination of the definition (2.23) for c^n_{n+1} reveals that it may be more succinctly stated as

$$c^n_{n+1} = \min\{x_n, w_{n+1}\} \quad . \tag{2.24}$$

Similar reasoning to that employed in calculating c^n_{n+1} applies to the computation of c^n_m. The removal of customer n either causes customers m and m-1 to be in the same busy period, or breaks the busy period after customer (m-1). The removal of customer n causes customer m-1 to save system time c^n_{m-1}. Hence, customer m-1 departs at an earlier time $d_{n-1} - c^n_{m-1}$. If this new departure time for customer m-1 is greater than the arrival time a_m of customer m, customers m and m-1 remain in the same busy interval and customer m saves a system time c^n_{m-1}. However, if $a_m > d_{m-1} - c^n_{m-1}$, customer m begins a new busy period and saves its waiting time w_m. These relationships are summarized in the following rule for computing c^n_m.

$$c^n_m = \begin{cases} c^n_{m-1} & \text{for } d_{m-1} - c^n_{m-1} \geq a_m \\ d_{m-1} - a_m = w_m & \text{for } a_m > d_{m-1} - c^n_{m-1} \end{cases} \tag{2.25}$$

This rule may be expressed more compactly as

$$c^n_m = \min\{c^n_{m-1}, w_m\} \quad . \tag{2.26}$$

Hence, the algorithm for computing the c^n_m's may be summarized as

$$c_n^n = s_n$$

$$c_{n+1}^n = \min\{x_n, w_{n+1}\} \tag{2.27}$$

$$c_m^n = \min\{c_{m-1}^n, w_m\} \quad m = n+2, \ldots M \quad ,$$

where M is the number of customers served in the given busy period. The customer-removal algorithm is now completely specified by relations (2.21), (2.22), and (2.27). The N busy interval delay gradient estimator may be expressed as

$$\hat{D}'_{(N)} = \frac{1}{M_{(N)}} \sum_{i=1}^{N} \sum_{n=1}^{M_i} \sum_{m=n}^{M_i} c_{m,i}^n \quad , \tag{2.28}$$

where

$$M_{(N)} = \sum_{i=1}^{N} M_i \quad . \tag{2.29}$$

We may write relation (2.28) recursively as

$$\hat{D}'_{(k)} = \left(\frac{M_{(k-1)}}{M_{(k)}}\right) \hat{D}'_{(k-1)} + \Delta_k \quad , \tag{2.30}$$

where

$$\Delta_k \triangleq \sum_{n=1}^{M_k} \sum_{m=n}^{M_k} c_{m,k}^n \quad . \tag{2.31}$$

In our third algorithm, we simulate an increase in rate $\delta\lambda$ by a linear

contraction in time scale. Assume that $t_{j,i}$ denotes the arrival time of the j-th customer in the i-th busy period relative to the beginning of the observation interval. We define a new set of shifted arrival times by

$$t^*_{j,i} = (1 - \frac{\delta\lambda}{\lambda})t_{j,i} \quad . \tag{2.32}$$

Since $\delta\lambda$ represents an infinitesimal change in rate, we can choose it sufficiently small so none of the busy periods are shifted onto other busy periods by the time contraction.

We now consider the increment in system time that comes from each customer arriving a little earlier and hence, waiting a little longer. The waiting time of the j-th customer in the i-th busy period is defined by

$$w_{j,i} = \sum_{\ell=1}^{j-1} x_{\ell,i} - (t_{j,i} - t_{j,1}) \ , \tag{2.33}$$

where $x_{\ell,i}$ denotes the service requirement of the ℓ-th customer in the i-th busy period. If we substitute the shifted arrival times given by (2.32) into (2.33), we can relate the new waiting times $w^*_{j,i}$ to $w_{j,i}$ by

$$w^*_{j,i} = w_{j,i} + \frac{\delta\lambda}{\lambda}[t_{j,i} - t_{1,i}] \tag{2.34}$$

We define $\tau_{j,i}$ as the arrival time of the j-th customer in the i-th busy period relative to the start of that busy period. Hence, the additional system time over N busy periods that follows from (2.34) can be expressed as

$$\frac{\delta\lambda}{\lambda} \sum_{i=1}^{N} Z_i \ , \tag{2.35}$$

where

$$Z_i = \begin{cases} 0 & M_i = 1 \\ \sum_{\ell=2}^{M_i} \tau_{\ell,i} & M_i > 1 \end{cases} \tag{2.36}$$

A second contribution to the increment in system time follows from the fact that our time contraction procedure shifts the right edge of the observation interval, leaving a gap of duration $\frac{\delta\lambda}{\lambda}T_E$ in which additional customers could arrive. Let $D_c(\lambda)$ denote the average system time per customer. Then, $\lambda D_c(\lambda)$ represents the average total delay accumulated per unit time. Hence, the average increment in system time associated with an interval of length $\frac{\delta\lambda}{\lambda}T_E$ would be given as

$$(\frac{\delta\lambda}{\lambda}T_E)(\lambda D_c(\lambda)) = \delta\lambda T_E D_c(\lambda) \ . \tag{2.37}$$

Since both λ and $D_c(\lambda)$ are unknowns, we use the fact that asymptotically,

$$\lambda \sim \frac{\sum_{i=1}^{N} M_i}{T_E} \ , \tag{2.38}$$

and

$$D_c(\lambda) \sim \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} S_{j,i}}{\sum_{i=1}^{N} M_i} \tag{2.39}$$

where $S_{j,i}$ denotes the system time of the j-th customer in the i-th busy period. Combining the contributions to the incremental system time given by (2.35) and (2.37) with relations (2.38)-(2.39) we obtain the following N-busy period delay gradient estimator:

$$\hat{D}'_{(N)} = \frac{\left\{\left(\sum_{i=1}^{N} \sum_{j=1}^{M_i} S_{j,i}\right) + Z_i\right\}}{M_{(N)}} \tag{2.40}$$

where Z_i is defined by relation (2.36) and $M_{(N)}$ by (2.29). Relation (2.40) may be written in recursive form as

$$\hat{D}'_{(k)} = \frac{M_{(k-1)}}{M(k)} \hat{D}'_{(k-1)} + \Delta_k \quad , \tag{2.41}$$

where

$$\Delta_k \triangleq \sum_{\ell=1}^{M_k} S_{\ell,k} + Z_k \quad . \tag{2.42}$$

Alternatively, by employing the definition of $S_{\ell,k}$ and Z_k, we may express Δ_k as

$$\Delta_k = \sum_{\ell=1}^{M_k} (M_k - \ell + 1) x_{\ell,k} \quad . \tag{2.43}$$

3. STATISTICAL ANALYSIS OF THE ALGORITHMS FOR M/D/1 AND M/G/1 QUEUES

Having motivated and defined all three algorithms for estimating the delay gradient, we now proceed to analyze their asymptotic behavior as the observation interval and hence the number of busy periods N included

become large, for the special cases of some simple queues. We sketch the analysis of the asymptotic bias for all three algorithms in the case of an M/D/1 queue. In addition, for an M/D/1 queue we derive Cramer-Rao lower bounds for the r.m.s. error of any unbiased estimator using the same information as each of the three algorithms. Finally, we will discuss briefly the analysis of the asymptotic bias of the customer-removal and time-contraction algorithms in the case of M/G/1 queues.

For a general queue, the asymptotic bias b associated with each algorithm is formulated as

$$b = \lim_{N\to\infty} E\{\hat{D}'_{(N)}\} - \frac{\partial D}{\partial \lambda}, \tag{3.1}$$

where $D(\lambda)$ is the average total delay accumulated per unit time. If $D_c(\lambda)$ denotes the average delay per customer, then $D(\lambda)$ is expressed as $\lambda D_c(\lambda)$, and hence $\frac{\partial D}{\partial \lambda}$ may be written as

$$\frac{\partial D}{\partial \lambda} = D_c(\lambda) + \lambda \frac{\partial D_c(\lambda)}{\partial \lambda} . \tag{3.2}$$

Now the average system time per customer $D_c(\lambda)$ is expressible as

$$D_c(\lambda) = \bar{x} + \bar{w}(\lambda), \tag{3.3}$$

where $\bar{x}$ denotes the average service time and $\bar{w}(\lambda)$ denotes the average waiting time. Hence, $\frac{\partial D}{\partial \lambda}$ can be reformulated as a function of the average service requirement and waiting time as

$$\frac{\partial D}{\partial \lambda} = \bar{x} + \bar{w}(\lambda) + \lambda \frac{\partial \bar{w}(\lambda)}{\partial \lambda} \quad . \tag{3.4}$$

We can evaluate the above expression for all queues for which an explicit form of the waiting time distribution is available. For an M/G/1 queue, $\bar{w}(\lambda)$ is given in [7] as

$$\bar{w}(\lambda) = \bar{x} \frac{\rho(1 + c_b^2)}{2(1 - \rho)} \quad , \tag{3.5}$$

where $$\rho = \lambda \bar{x} \quad , \tag{3.6}$$

is the utilization factor and

$$c_b^2 = \frac{\sigma_b^2}{\bar{x}^2} \, , \tag{3.7}$$

where σ_b^2 denotes the variance of the service time distribution.

We now briefly describe the steps employed to show the asymptotic unbiasedness of the customer-addition algorithm in the case of an M/D/1 queue. For the full details we refer the reader to [8]. From relation (3.1) we must evaluate

$$\lim_{N \to \infty} E \, \hat{D}'_{(N)} = \lim_{N \to \infty} E\{\frac{1}{T_E} \frac{\delta \hat{S}_{(N)}}{\delta \lambda}\} \quad , \tag{3.8}$$

where $\delta \hat{S}_{(N)}$ is defined by relation (2.5), in order to compute the asymptotic bias. Letting $\bar{M}$ denote the average number of customers per busy period, by the law of large numbers we have

$$\lim_{N\to\infty} T_E = \lim_{N\to\infty}\{\frac{N\bar{M}}{\lambda}\} \ , \tag{3.9}$$

and interchanging the limit and expectation operations in (3.8) we need to compute

$$\lambda \lim_{N\to\infty} \{\frac{1}{N\bar{M}} E\{\frac{\delta\hat{S}_{(N)}}{\delta\lambda}\} \} \ . \tag{3.10}$$

Now define the two aggregate vectors

$$\tilde{M}_{(N)} = (M_1,\ldots,M_N) \ , \tag{3.11}$$

and

$$\tilde{I}_{(N)} = (I_1,\ldots,I_{N-1}) \ , \tag{3.12}$$

where the M_i's are the number of customers served in the i-th busy interval and the I_i's are the i-th idle period durations. Then for an M/D/1 queue, the queueing record is completely specified by $\tilde{M}_{(N)}$ and $\tilde{I}_{(N)}$. We break up the calculation of $E\{\frac{\delta\hat{S}_{(N)}}{\delta\lambda}\}$ by first conditioning on $\tilde{M}_{(N)}$, averaging over $\tilde{I}_{(N)}$, and performing a final averaging over $\tilde{M}_{(N)}$. Writing out $E\{\frac{\delta\hat{S}_{(N)}}{\delta\lambda}\}$ from relation (2.5) as

$$E\{\frac{\delta\hat{S}_{(N)}}{\delta\lambda}\} = \sum_{k=1}^{N} E\{E(\delta S_{(N)} \mid t\in T_k, \tilde{M}_{(N)}, \tilde{I}_{(N)})\} + \sum_{k=1}^{N} E\{E(\delta S_{(N)} \mid t\in I_k \tilde{M}_{(N)}, \tilde{I}_{(N)}\} \tag{3.13}$$

We note that there are two types of terms that must be analyzed, corresponding to the busy and idle period contributions to the incremental delay. To

perform the initial averaging of $\frac{\delta\hat{S}_{(N)}}{\delta\lambda}$ over $\tilde{I}_{(N)}$ we use our expressions for $E(\delta S_{(N)} | t\epsilon T_k, \tilde{M}_{(N)}, \tilde{I}_{(N)})$ and $E(\delta S_{(N)} | t\epsilon I_k, \tilde{M}_{(N)}, \tilde{I}_{(N)})$ given by relations (2.10), (2.12) and the fact that the I_j's are independent exponentially distributed random variables with probability densities

$$p_{I_j}(\tau) = \lambda e^{-\lambda\tau} u_{-1}(\tau) \quad . \tag{3.14}$$

After considerable manipulations, which are detailed in [8], we obtain the mean of $\frac{\delta\hat{S}_{(N)}}{\delta\lambda}$ conditioned on $\tilde{M}_{(N)}$ as

$$\begin{aligned} E\{\frac{\delta\hat{S}_{(N)}}{\delta\lambda} | \tilde{M}_{(N)}\} = & \sum_{k=1}^{N} (M_k x^2 + \frac{1}{2}M_k^2 x^2) + \sum_{j=1}^{N-1}\sum_{k=1}^{N-j} M_j M_{j+k} a_k \\ & + \sum_{k=1}^{N-1} (\frac{x}{\lambda} + \frac{1}{\lambda}M_{k+1}(x + \frac{1}{\lambda}(e^{-\lambda x} - 1))) \\ & + \sum_{k=1}^{N-2}\sum_{j=2}^{N-k} M_{k+j} b_j \\ & + \sum_{k=1}^{N-2}\sum_{j=2}^{N-k} M_{k+j} c_j \end{aligned} \tag{3.15}$$

where

$$a_k = x(x - \frac{k}{\lambda})e^{-\lambda x}\sum_{\ell=k}^{\lambda}\frac{(\lambda x)^\ell}{\ell!} + x^2 e^{-\lambda x}\frac{(\lambda x)^{k-1}}{(k-1)!} , \tag{3.16}$$

$$b_j = e^{-\lambda x}\{\frac{x}{\lambda}\sum_{\ell=j-1}^{\infty}\frac{(\lambda x)^{\ell+2}}{(\ell+2)!} - \frac{j}{\lambda^2}\sum_{\ell=j-1}^{\infty}\frac{(\lambda x)^{\ell+3}}{(\ell+3)!}\}, \tag{3.17}$$

and

$$c_j = \frac{(\lambda x)^j}{(j+1)!}e^{-\lambda x}\frac{x}{\lambda} \quad . \tag{3.18}$$

We now note that the M_i's are independent, identically distributed random variables and that for an M/D/1 queue the first and second moments $\bar{M}$ and $\overline{M^2}$ are given by

$$\bar{M} = \frac{1}{1-\rho} , \tag{3.19}$$

and

$$\overline{M^2} = \frac{2\rho - \rho^2}{(1-\rho)^3} + \frac{1}{(1-\rho)} , \tag{3.20}$$

where

$$\rho = \lambda\bar{x} . \tag{3.21}$$

By using (3.19) and (3.20) we can compute the unconditional expectation of $\frac{\delta S_{(N)}}{\delta\lambda}$. Then by employing (3.15)-(3.18) we are able to evaluate the limit

$$\lim_{N\to\infty} \frac{\lambda}{N\bar{M}} E\{\frac{\delta\hat{S}_{(N)}}{\delta\lambda} = x + \frac{\rho x}{2(1-\rho)} + \frac{\rho x}{2(1-\rho)^2} . \tag{3.22}$$

The above result (3.22) is exactly the quantity obtained by specializing the expression (3.5) for $\bar{w}(\lambda)$ to the case of an M/D/1 queue and substituting into our general expression (3.4) for the delay gradient $-\frac{\partial D}{\partial\lambda}$. Hence, the customer-addition algorithm is asymptotically unbiased for an M/D/1 queue.

Since the calculation of the exact variance associated with the customer-addition algorithm is too cumbersome, we derive a Cramer-Rao bound. If we have an observation vector R, a parameter α we want to estimate and a density for R parameterized by α - $p(R:\alpha)$, the Cramer-Rao lower bound for

the variance of any unbiased estimator $\hat{\alpha}(R)$ of α is stated as follows:

$$\mathrm{Var}(\hat{\alpha}(R) - \alpha) \geq \frac{1}{-E\{\frac{\partial^2 \ln p(R:\alpha)}{\partial \alpha^2}\}} \quad . \tag{3.23}$$

In our case we identify R as $(\tilde{M}_{(N)}, \tilde{I}_{(N)})$. As a product of the "memoryless" property associated with the exponential interarrival time distribution we can show that the M_i's and I_j's are all mutually independent so that $p(\tilde{M}_{(N)}, \tilde{I}_{(N)})$ is given as

$$p(\tilde{M}_{(N)}, \tilde{I}_{(N)}) = \prod_{i=1}^{N} \Pr\{M_i = m_i\} \prod_{j=1}^{N-1} \lambda e^{-\lambda I_j} \quad , \tag{3.24}$$

where for an M/D/1 queue $\Pr\{M_i = m_i\}$ is given in [7] as

$$\Pr\{M_i = m_i\} = \frac{(m_i \rho)^{m_i - 1}}{m_i!} e^{-m_i \rho} \quad . \tag{3.25}$$

Hence, pretending that ρ is the parameter of interest, we can express $p(\tilde{M}_{(N)}, \tilde{I}_{(N)}:\rho)$ as

$$p(\tilde{M}_{(N)}, \tilde{I}_{(N)}:\rho) = \left(\prod_{i=1}^{N} \frac{m_i^{m_i - 1}}{m_i!} \right) \left(\rho^{\sum_{i=1}^{N} m_i - N} \right) e^{-\sum_{i=1}^{N} m_i \rho} \left(\frac{\rho}{x} \right)^{N-1} e^{-\frac{\rho}{x} \sum_{j=1}^{N-1} I_j} \tag{3.26}$$

But now letting γ denote $\frac{\partial D}{\partial \lambda}$ for an M/D/1 queue, where $\frac{\partial D}{\partial \lambda}$ is defined in relation (3.4), we can express ρ as a function of γ as follows:

$$\rho = 1 - \sqrt{1 - \frac{2(\gamma - x)}{(2\gamma - x)}} \quad . \tag{3.27}$$

Hence, we can form $p(\tilde{M}_{(N)}, \tilde{I}_{(N)} : \gamma)$ and can evaluate the bound in relation (3.23) as

$$\mathrm{Var}(\gamma - \hat{\gamma}) \geq \frac{x^2 \rho^2}{(N-1+\rho)(1-\rho)^5} \tag{3.28}$$

Next we investigate the asymptotic properties of the customer removal algorithm in the case of an M/G/1 queue by first interpreting the terms in the estimator. For a given busy period, the inner two summations in relation (2.28) may be grouped into two terms representing the sum of all the service times of customers in that busy period and the cumulative service time saved by by all other customers due to the removal of each customer separately. Hence, the N busy period delay gradient estimator may be expressed as

$$\hat{D}'_{(N)} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} S_{j,i}}{M_{(N)}} + \frac{\sum_{i=1}^{N} p_i}{M_{(N)}} , \tag{3.29}$$

where

$$p_i = \begin{cases} 0 & M_i = 1 \\ \sum_{n=1}^{M_i - 1} \sum_{m=n+1}^{M_i} c^n_{m,i} & M_i > 1 \end{cases} . \tag{3.30}$$

We examine the asymptotic behavior of the mean of the estimator specified in (3.29) by interchanging the expectation and limit operation. By appealing to the law of large numbers the limiting form of the estimator as N becomes unbounded is

$$\lim_{N\to\infty} E\,\hat{D}'_{(N)} = E \lim_{N\to\infty} \hat{D}'_{(N)} = D_c(\lambda) + \frac{\bar{p}}{\bar{M}}, \tag{3.31}$$

where $D_c(\lambda)$ is the average system time per customer, $\bar{p}$ is the expectation of the quantity defined in (3.30), and $\bar{M}$ is the expected number of customers served per busy period. Employing the general relation (3.4) which expresses the delay gradient as equal to $D_c(\lambda)$ plus $\frac{\partial \bar{w}(\lambda)}{\partial \lambda}$, we can formulate the asymptotic bias of the customer-removal algorithm as

$$b = \lim_{N\to\infty} E\,\hat{D}'_{(N)} - \frac{\partial D}{\partial \lambda} = \frac{\bar{p}}{\bar{M}} - \lambda \frac{\partial \bar{w}(\lambda)}{\partial \lambda}. \tag{3.32}$$

We can break up the calculation of $\bar{p}$ by conditioning on $M = i$ for $i = 2,\ldots,\infty$ and hence we find

$$\bar{p} = \sum_{i=2}^{\infty} Q(i) f_i, \tag{3.33}$$

where f_i is the probability that i customers are served in a busy period and $Q(i)$ is defined as

$$Q(i) = \sum_{n=1}^{i-1} \sum_{m=n+1}^{i} E(c_m^n \mid M=i). \tag{3.34}$$

Now for an M/G/1 queue the average number of customers served per busy period is given in [7] by

$$\bar{M} = \frac{1}{1-\rho}. \tag{3.35}$$

The z-transform for the probability density of the number of customers served per busy period is described in [7] by the following functional equation:

$$F(z) = zB^*(\lambda - \lambda F(z)), \tag{3.36}$$

where $B^*(s)$ is the one-sided Laplace transform of the service time density and $F(z)$ is defined by

$$F(z) = \sum_{n=1}^{\infty} f_n z^n . \tag{3.37}$$

We may now employ the expression for $\bar{w}(\lambda)$ in (3.5) to write $\lambda \frac{\partial w(\lambda)}{\partial \lambda}$ as

$$\lambda \frac{\partial \bar{w}(\lambda)}{\partial \lambda} = (1-\rho) \frac{\bar{x}}{2} (1 + c_b^2) \sum_{k=0}^{\infty} \frac{(k+2)(k+1)}{2} \rho^{k+1} , \tag{3.38}$$

where $\rho = \lambda \bar{x}$ is the utilization factor. Hence, for an M/G/1 queue the asymptotic bias may be expressed as

$$b = (1-\rho)[\sum_{n=2}^{\infty} Q(n) f_n - \frac{\bar{x}(1+c_b^2)}{4} \sum_{k=0}^{\infty} (k+2)(k+1)\rho^{k+1}] . \tag{3.39}$$

When $\bar{p}$ may be expressed as a power series in ρ, relation (3.39) expresses b as an expansion in powers of ρ.

We now comment briefly on the calculation of the Q(i)'s in relation (3.34). A complete discussion of this problem is contained in [8]. To calculate Q(i), we need to evaluate conditional expectations of the form:

$$E\{c_{n+k}^n | M=i\} , \tag{3.40}$$

for n = 2...M, k=1..M-n. From relations (2.27), we can express c_{n+k}^n as

$$c_{n+k}^n = \min\{x_n, w_{n+1}, w_{n+2} \cdots w_{n+k}\} \tag{3.41}$$

Our approach is to derive the density for $(x_n, w_{n+1}..w_{n+k})$ conditioned on $M = i$ and then compute $F_{c^n_{n+k}|M=i}(\tau)$, the distribution function for c^n_{n+k} given $M = i$ as

$$F_{c^n_{n+k}|M=i}(\tau) = 1 - \Pr\{x_n > \tau,\ w_{n+1} > \tau,\ w_{n+k} > \tau\} . \tag{3.42}$$

Now we calculate $P_{c^n_{n+k}|M=i}(\tau)$ as

$$P_{c^n_{n+k}|M=i}(\tau) = \frac{d}{d\tau} F_{c^n_{n+k}|M=i}(\tau) , \tag{3.43}$$

and derive the conditional mean (3.40) as

$$E(c^n_{n+k}|M=i) = \int \tau p_{c^n_{n+k}|M=i}(\tau)\, d\tau . \tag{3.44}$$

The calculations become simpler in the case of an M/D/1 queue where

$$c^n_{n+k} = \min\{x,\ z_{n,k}\} , \tag{3.45}$$

with

$$z_{n,k} \triangleq \min\{w_{n+1},\ldots,w_{n+k}\} . \tag{3.46}$$

Here we calculate the joint density of $(w_{n+1}..w_{n+k})$ conditional on $M = i$ and then compute $F_{z_{n,k}|M=i}(\tau)$, the distribution function for $z_{n,k}$ given $M = i$ as

$$F_{z_{n,k}|M=i}(\tau) = 1 - \Pr\{w_{n+1} > \tau,..w_{n+k} > \tau\} . \tag{3.47}$$

Given the distribution function for $z_{n,k}$ conditioned on M = i we compute the density for $z_{n,k}$ conditioned on M = i by

$$P_{z_{n,k}|M=i}(\tau) = \frac{d}{d\tau} F_{z_{n,k}|M=i}(\tau) \quad , \tag{3.48}$$

and can write down the conditional density for c^n_{n+k} explicitly as

$$P_{c^n_{n+k}|M=i}(\tau) = (1 - F_{z_{n,k}|M=i}(x))\,\delta(\tau - x) + P_{z_{n,k}|M=i}(\tau)\, U_{-1}(x - \tau) \tag{3.49}$$

By employing the representation for the asymptotic bias (3.39), the technique for computing the Q(i)'s sketched in the preceding paragraphs, and known formulae for the f_i's we were able to show that for an M/D/1 queue, the asymptotic bias b may only contain terms of fourth order or higher in ρ. We suspect strongly however that the coefficients of all powers of ρ in the expansion of b obtained from (3.39) vanish for an M/D/1 queue, although we were not able to prove this explicitly. Analysis of relation (3.39) for the case of an M/M/1 queue reveals that b contains nonzero terms of order ρ and higher.

Having analyzed the asymptotic bias associated with the customer-removal procedure, we now proceed to derive a Cramer-Rao bound for the variance of any unbiased estimator of the delay gradient that works with the same observations as the customer-removal algorithm in the case of an M/D/1 queue. Since for an M/D/1 queue the service requirement is a deterministic quantity - x, the observations which the customer removal algorithm employs are the M_i's, the number of customers served in the i-th busy period, and the $w_{j,i}$'s, the waiting times of the $j = 2,\ldots,M_i$ customers

in the i-th busy period. Hence, we define the N busy period observation vector as

$$Y_{(N)} \triangleq (Y_1 \vdots \cdots \vdots Y_N) \ , \tag{3.50}$$

where $$Y_i \triangleq (M_i, w_{2,i} \cdots w_{M_i,i}) \ . \tag{3.51}$$

Since waiting times and the number of customers served per busy period are statistically independent from one busy period to the next, the joint density $P(Y_{(N)})$ may be expressed as

$$P(Y_{(N)}) = \prod_{i=1}^{N} P(Y_i) \tag{3.52}$$

where we can decompose $P(Y_i)$ as

$$P(Y_i) = P(M_i) P(w_{2,i} \cdots w_{M_i,i} \mid M_i) \ . \tag{3.53}$$

Employing the expression for $p(M_i)$ given by (3.25) and the expression for $p(w_{2,i} \cdots w_{M_i,i} \mid M_i)$ derived in [8], we can form $P(Y_{(N)}:\rho)$ as

$$p(Y_{(N)}:\rho) = \frac{\rho^{(\sum_{i=1}^{N} M_i - N)}}{x^{(\sum_{i=1}^{N} M_i - N)}} e^{(-\sum_{i=1}^{N} M_i)\rho} \ , \tag{3.54}$$

where $$0 \leq w_{2,i} \leq x \ , \tag{3.55}$$

and $$0 \leq w_{k+1,i} \leq x + w_{k,i} \quad k = 2..M_i - 1 \tag{3.56}$$

Now, using the relation between the utilization factor ρ and the delay gradient γ, for an M/D/1 queue, given by relation (3.27), we may form $p(Y_{(N)}:\gamma)$ and compute the Cramer-Rao band from (3.23) as

$$\mathrm{Var}(\gamma - \hat{\gamma})^2 \geq \frac{x^2 \rho}{N(1-\rho)^5} \quad . \tag{3.57}$$

Finally, we now determine the asymptotic behavior of the time-contraction procedure. From relation (2.40), we find by exchanging limit and expectation operations and appealing to the law of large numbers that

$$\lim_{N\to\infty} E\hat{D}'_{(N)} = D_c(\lambda) + \frac{\bar{z}}{\bar{M}} \quad , \tag{3.58}$$

where $\bar{z}$ denotes the expected value of z_i defined by relation (2.36). Hence, by the same reasoning as for the customer-removal algorithm, the asymptotic bias b is given as

$$b = \frac{\bar{z}}{\bar{M}} - \lambda \frac{\partial w(\lambda)}{\partial \lambda} \quad . \tag{3.59}$$

Similarly, we may compute $\bar{z}$ by conditioning on $M = i$, $i = 2..\infty$ as

$$\bar{z} = \sum_{i=2}^{\infty} \Omega(i) f_i \quad , \tag{3.60}$$

where f_i denotes the probability of i customers being served in the i-th busy period and $\Omega(i)$ is defined as

$$\Omega(i) = E\left(\sum_{i=2}^{M} \tau_i \middle| M = i \right) \tag{3.61}$$

If we let

$$\tilde{\tau}_j \triangleq \sum_{\ell=2}^{j} \tau_i, \tag{3.62}$$

we show in [8] that it is possible to derive the joint density of the $\tilde{\tau}_j$'s for j = 2..M conditioned on M = i and hence express the desired conditional expectation (3.61) as a single integral. Thus, in the same manner as for the customer-removal algorithm, relation (3.59) becomes a representation in powers of ρ for the asymptotic bias in the case of M/G/1 queues. By employing relation (3.38) for $\lambda\frac{\partial \bar{w}(\lambda)}{\partial \lambda}$ we obtain

$$b = (1 - \rho)\left(\sum_{i=2}^{\infty} \Omega(i) f_i - \frac{\bar{x}(1 + c_b^2)}{4} \sum_{k=0}^{\infty} (k+2)(k+1)\rho^{k+1}\right) \tag{3.63}$$

By employing relation (3.63), we showed that for an M/M/1 queue the asymptotic bias for the time-contraction algorithm contains nonzero terms of first order in ρ and higher. By analyzing the form of the time-contraction estimator obtained by using relation (2.43) to define Δ_k, we are able to show asymptotic unbiasedness of the algorithm for an M/D/1 queue. From relation (2.43), we may alternatively express the N busy period estimator as

$$\hat{D}'_{(N)} = \frac{\sum_{j=1}^{N} B_j}{M_{(N)}}, \tag{3.64}$$

where

$$B_j = \sum_{\ell=1}^{M_j} (M_j - \ell + 1) x_{\ell,j} \quad . \tag{3.65}$$

Hence for an M/D/1 queue, $\hat{D}'_{(N)}$ would be given by

$$\hat{D}'_{(N)} = \frac{\sum_{j=1}^{N} \frac{1}{2}(M_j^2 + M_j) x}{M_{(N)}} \quad . \tag{3.66}$$

Now, by exchanging limit and expectation operations and appealing to the law of large numbers we can show that

$$\lim_{N\to\infty} \hat{D}'_{(N)} = \frac{1}{2} x \{ \frac{\overline{M^2}}{\bar{M}} + 1 \} \quad . \tag{3.67}$$

Employing earlier (3.19) - (3.20) expressions for $\overline{M^2}$ and $\bar{M}$, we can verify that the limit in (3.67) is exactly $\frac{\partial D}{\partial \lambda}$ for an M/D/1 queue.

We now conclude our analysis of the time-contraction algorithm by deriving a Cramer-Rao band on the variance of any unbiased estimator of the delay gradient that employs the same observations, in the case of an M/D/1 queue. Relation (3.66) indicates that for an M/D/1 queue the observation vector consists of $(M_1,\ldots,M_N)$ - the numbers of customers served in each busy period respectively. Since the number of customers served is independent from one busy period to another and for an M/D/1 queue the probability of M customers in a busy period is given by (3.25), $p(M_1..M_N:\rho)$ is given as follows:

$$p(M_1,..M_N:\rho) = \prod_{i=1}^{N} \frac{(M_i\rho)^{M_i-1}}{M_i!} e^{-M_i\rho} \quad . \tag{3.68}$$

Now, by employing relation (3.27), which expresses ρ as a function of the delay gradient γ, we may form $p(M_1..M_N:\gamma)$ and hence obtain the following lower band on the variance of any unbiased estimate $\hat{\gamma}$:

$$\mathrm{Var}(\gamma - \hat{\gamma}) \geq x^2 \frac{\rho}{N(1-\rho)^5} \quad . \tag{3.69}$$

4. COMPUTATIONAL COMPLEXITY AND STORAGE REQUIREMENT ANALYSIS FOR THE THREE ALGORITHMS

Having completed our examination of the statistical properties of all three estimation algorithms, we proceed to compare and contrast the computational complexity and storage requirements of each procedure. Since the estimate updating at the end of the k-th busy period assumes the form in relation (2.4) for each algorithm, we consider the storage and computational requirement of forming Δ_k.

For the customer addition algorithm, Δ_k is defined by relation (2.16). To compute Δ_{I_k} and Δ_{B_k} we need to have stored $(I_{n_k} \ldots I_{k-1})$ and $T_{n_{k+1}} \ldots T_k)$. Letting

$$n_{b,k} = k - n_k \quad , \tag{4.1}$$

we need $2n_{b,k}$ storage locations for the busy and idle period durations during which an additional arrival can cause customers of the most recent k-th busy period to suffer extra delay. Since $n_{b,k}$ is a random variable, in designing a system to implement the customer-addition procedure we would

need a rational for choosing a buffer size $N_b(\varepsilon)$, such that

$$\Pr\{n_{b,k} > N_b(\varepsilon)\} < \varepsilon \quad , \tag{4.2}$$

for some specified tolerable buffer overflow probability ε. For M/G/1 queues, a manner for choosing $N_b(\varepsilon)$ is discussed in [8]. The remaining component of Δ_k is Δw_k which is defined by relation (2.9). Thus, the storage requirement for forming Δ_k consists of $2n_{b,k} + 0(1)$ storage locations, where $0(1)$ denotes some constant. If we consider the computations involved in forming Δw_k, ΔI_k, and ΔB_k we count a total of

$$\begin{aligned} &7n_{b,k} + 0(1) && \text{additions,} \\ &6n_{b,k} + 0(1) && \text{multiplications,} \\ &2n_{b,k} + 0(1) && \text{divisions,} \\ \text{and} \quad &n_{b,k} && \text{comparisons.} \end{aligned} \tag{4.3}$$

For the customer removal algorithm Δ_k is defined by relation (2.31). By changing the order of the summation in (2.31), we can conceptualize Δ_k as being evaluated while the k-th busy interval progresses and requiring $M_k + 0(1)$ storage locations. The number of operations required are

$$\begin{aligned} &\frac{M_k(M_k - 1)}{2} \text{ comparisons,} \\ \text{and} \quad &\frac{M_k(M_k - 1)}{2} + 4M_k + 0(1) \text{ additions.} \end{aligned} \tag{4.4}$$

In the time-contraction algorithm Δ_k is defined by relation (2.42). Since Δ_k can be evaluated as the busy period progresses by simply adding service requirements, waiting times, and arrival times relative to the start of the busy interval, we only need 0(1) storage locations and $3M_k$ + 0(1) additions for its computation.

From the above discussion we see that of all three algorithms, the time-contraction procedure requires no buffer with randomly varying size. In addition, the computational load of the time-contraction algorithm is the least of the three. Hence, the time-contraction procedure is the least costly to implement.

5. SIMULATION RESULTS AND CONCLUSIONS

Earlier we analyzed the asymptotic bias of the three estimation algorithms. Since we are unable to calculate the variance of the estimators as a function of N, the question of consistency, whether the estimates converge asymptotically to the delay gradient, remains unanswered. In addition, for an M/D/1 queue we would like to know if our algorithms are asymptotically efficient, whether they achieve the Cramer-Rao bounds derived in (3.28), (3.57), and (3.69). We would also like to investigate the robustness of the customer-removal and time-contraction estimation procedures by seeing how they perform for a variety of queueing systems. We attempt to answer these questions by presenting the results of simulating all three algorithms for an M/D/1 queue and simulating the customer-removal and time-contraction algorithms for M/M/1, D/M/1, and U/M/1 queues.

We simulate a single-server queue by the following recursion for successive waiting times:

$$w_{n+1} = \max\{w_n + x_n - \theta_n, 0\} \text{ with } w_1 = 0 , \tag{5.1}$$

where x_n and θ_n are random variables corresponding to the n-th service requirement and the inter-arrival time between the n-th and (n+1)-st customer respectively. When w_ℓ goes to zero, this signals the start of a new busy period.

We now describe the calculations necessary to evaluate the statistics of a given estimator. To derive estimates for the mean and variance of the delay derivative estimate based on k busy periods, $\hat{D}'_{(k)}$, we generate a certain sample size N_s of k-busy period records, processing each to form a delay gradient estimate $\hat{D}'_{(k),i}$. We compute estimates of the bias $\hat{b}(k)$ and variance $\hat{\sigma}^2_{(k)}$ associated with $\hat{D}'_{(k)}$ as follows:

$$\hat{b}_{(k)} = \frac{1}{N_s} \sum_{i=1}^{N_s} \hat{D}'_{(k),i} - \frac{\partial D}{\partial \lambda} , \tag{5.2}$$

and

$$\hat{\sigma}^2_{(k)} = \frac{N_s}{N_s - 1} \left(\frac{1}{N_s} \sum_{i=1}^{N_s} \hat{D}'^2_{(k),i} - \left(\frac{1}{N_s} \sum_{i=1}^{N_s} \hat{D}'_{(k),i} \right)^2 \right) . \tag{5.3}$$

The measure of performance we are most interested in is the fractional r.m.s. error:

$$f(k) = \frac{\sqrt{E\left(\frac{\partial D}{\partial \lambda} - \hat{D}'_{(k)}\right)^2}}{\frac{\partial D}{\partial \lambda}} . \tag{5.4}$$

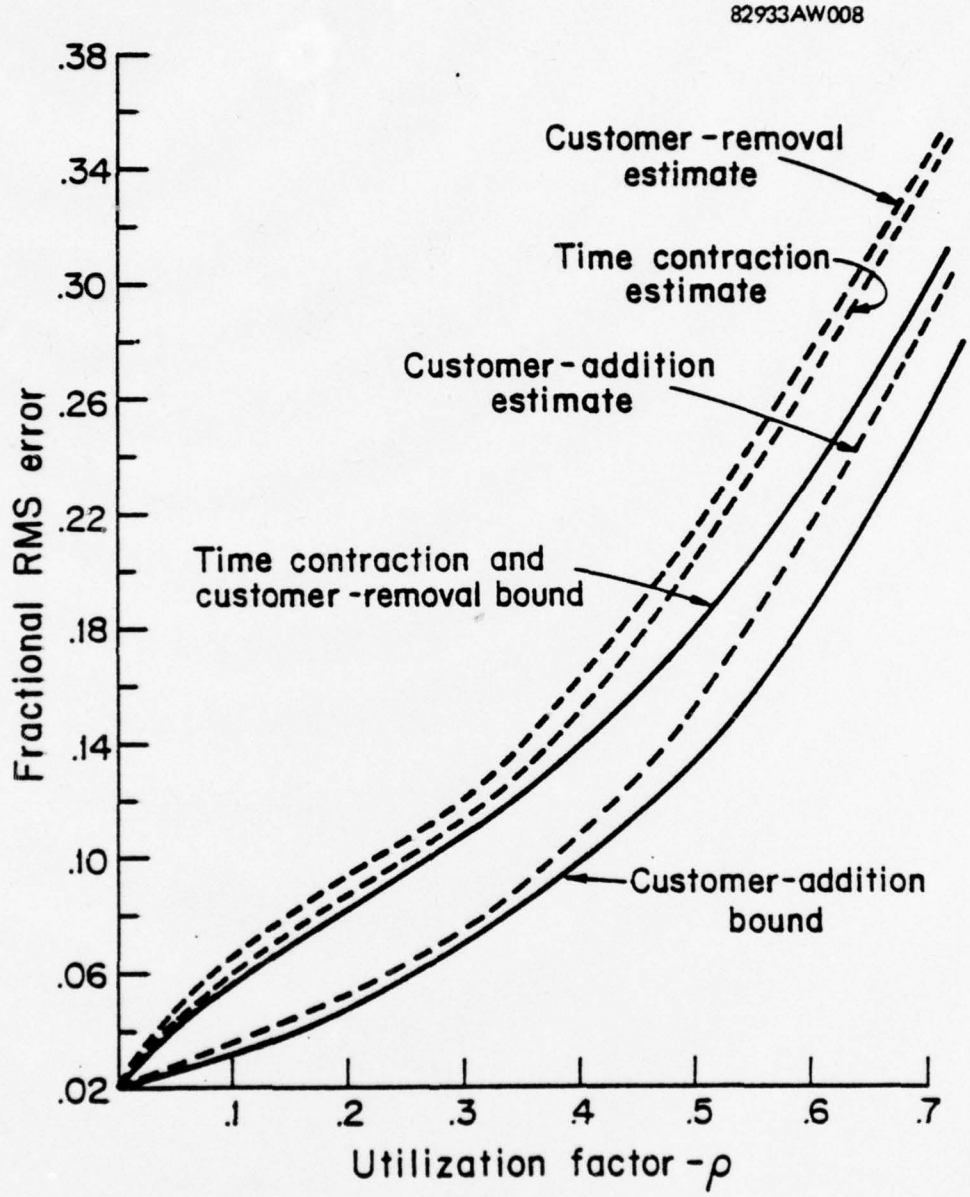
We approximate f(k) by $\hat{f}(k)$ defined as

$$\hat{f}(k) = \frac{\sqrt{\hat{b}^2(k) + (\gamma(k))^2}}{\frac{\partial D}{\partial \lambda}} . \tag{5.5}$$

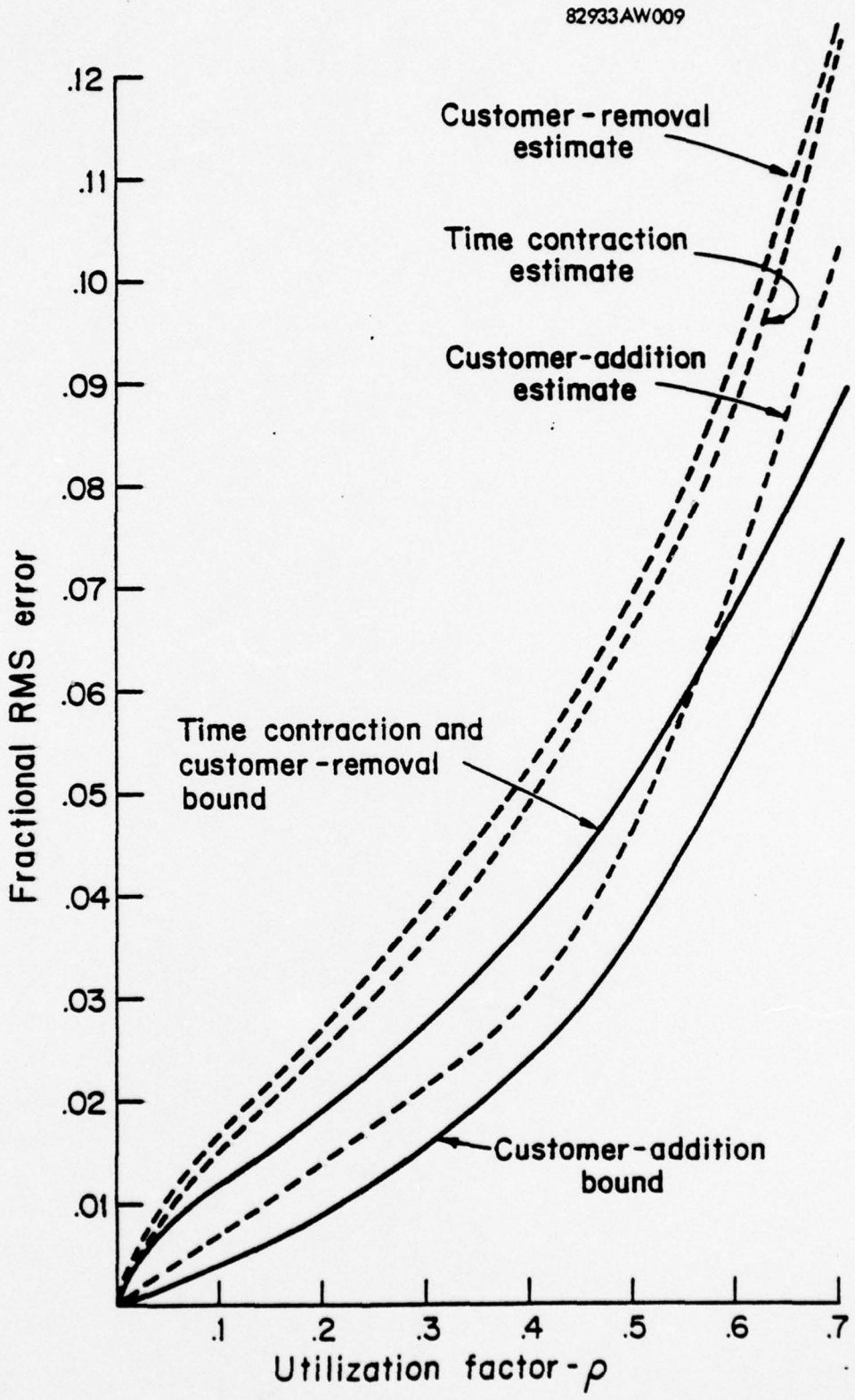
The value for N_s employed was arrived at by experimentation as $N_s = 400$. For each type of queue of interest we computed $\hat{f}(k)$ for k = 10, 100, and 1000 with the utilization factor $-\rho = \lambda\bar{x}$ - varying between .1 and .7.

Having defined the measure f(k) which we will use to compare the algorithms, we first discuss the simulation results for an M/D/1 queue. Curves of $\hat{f}(k)$ for k = 100 and 1000 are presented in graphs (5.1) and (5.2). The consistency of the three procedures is suggested by the improved performance with increasing N. Together with $\hat{f}(k)$, we display the fractional r.m.s. errors that follow from our Cramer-Rao bounds for the variance of each algorithm. The closeness of the simulation curves to their respective lower bounds suggest that all three algorithms are asymptotically efficient for an M/D/1 queue.

The bounds displayed in graphs (5.1) and (5.2) are reflections of the amount of information about the delay gradient contained in the observations which each algorithm employs. In actual practice, the rate of message arrivals to the queues associated with each link in a Computer-Network would be perturbed periodically by some global routing procedure. Assuming that the adjustments in arrival rates are small and that we have available an old estimate for the delay gradient, the number of busy periods of observations required to achieve a given r.m.s. estimation error may be significantly less than that predicted by graphs (5.1)-(5.2).



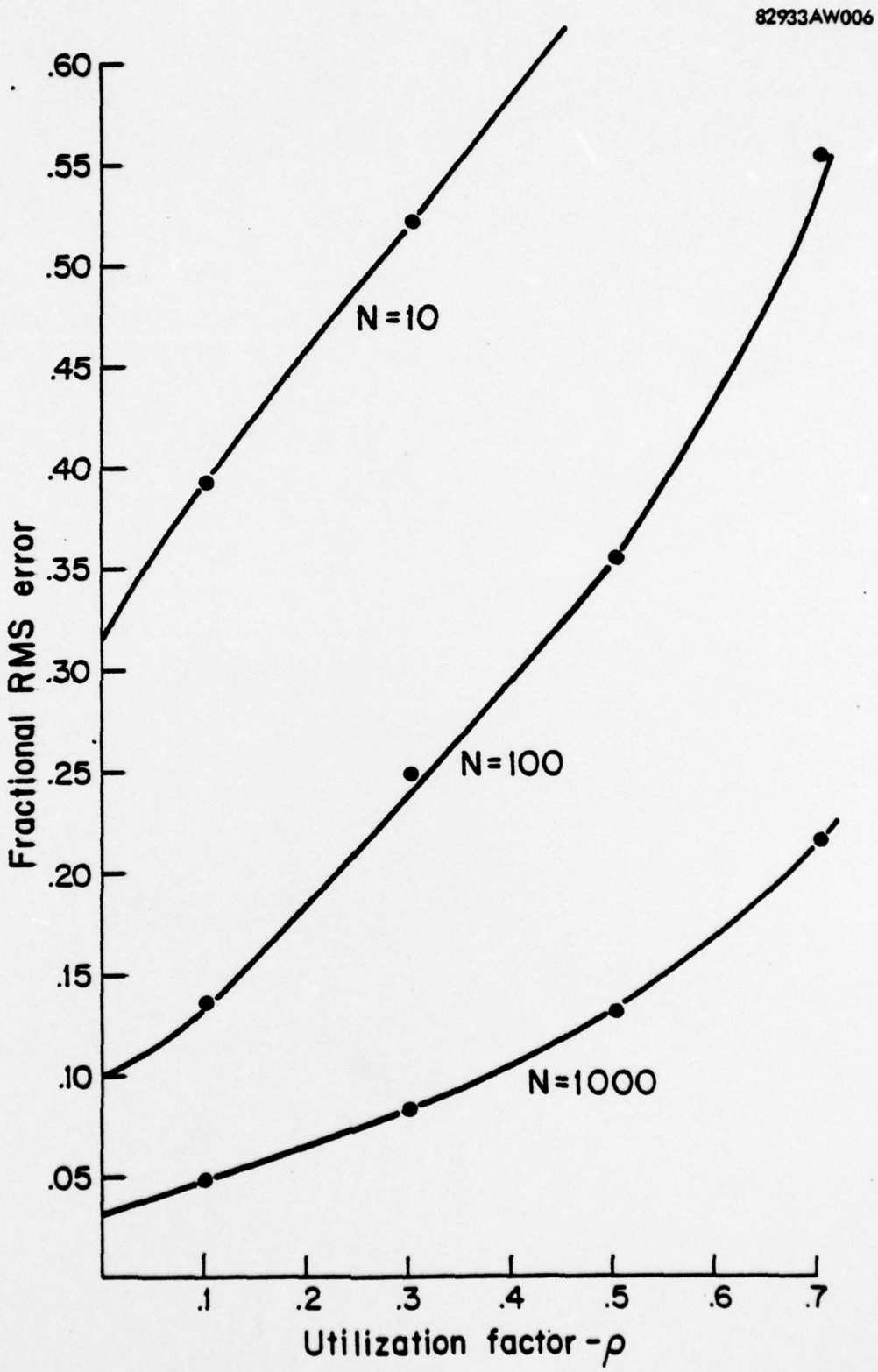
Graph (5.1) Lower Bounds on Fractional RMS Error for M|D|1(k=100).



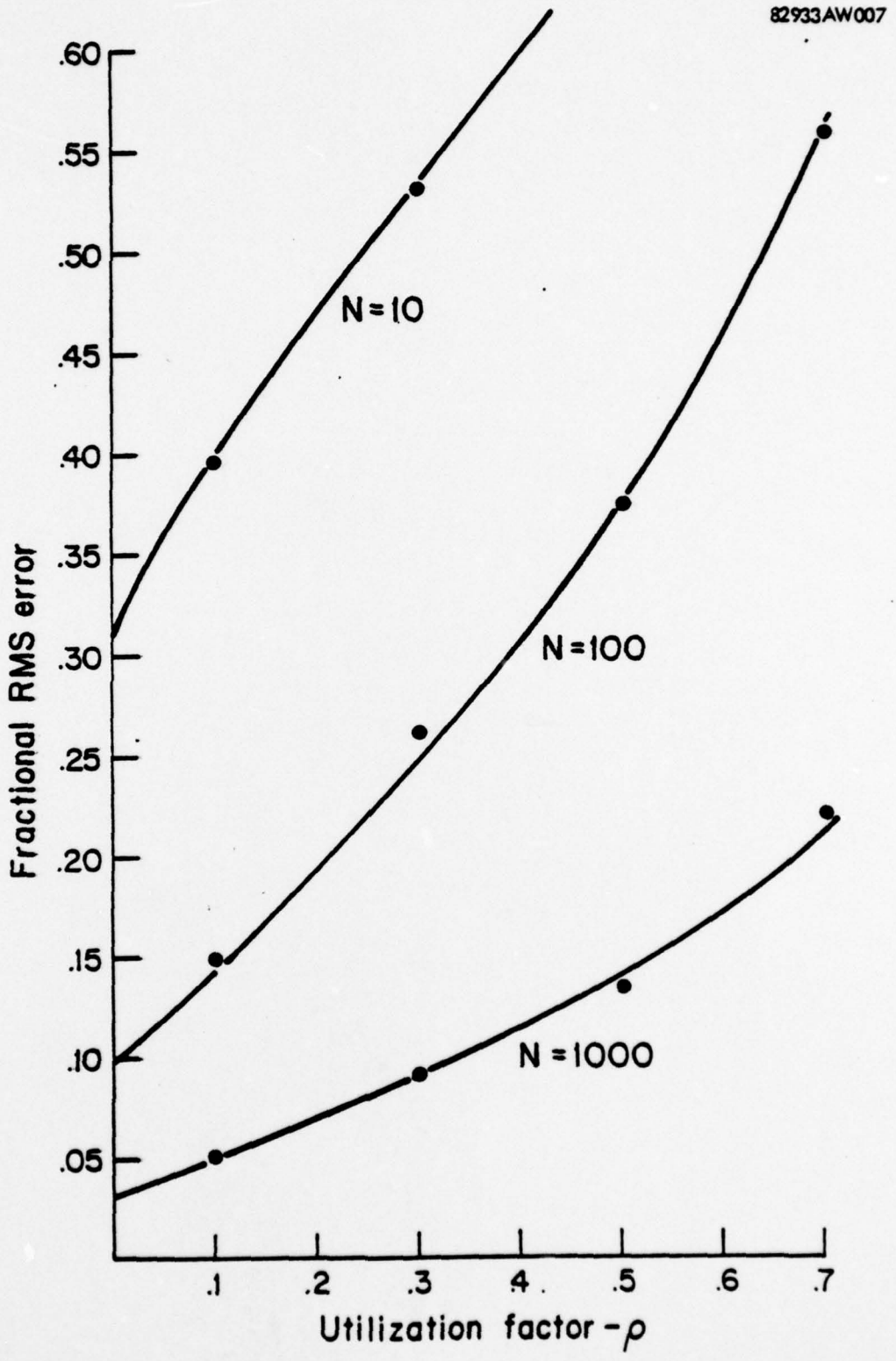
Graph (5.2) Lower Bounds on Fractional RMS Error for M|D|1. (k=1000).

Now we examine the robustness properties of the customer removal and time-contraction algorithms by comparing their performance for M/M/1, U/M/1, and D/M/1 queues. The simulation results for the fractional r.m.s. error of the two algorithms for M/M/1, U/M/1, and D/M/1 queues are presented in graphs (5.3)-(5.8), respectively. The time contraction algorithm performs slightly better than the customer-removal algorithm for M/M/1 and both perform nearly the same for a U/M/1 queue. The only dramatic differences occur for the D/M/1 queue, where the time-contraction procedure performs better than the customer-removal algorithm. This result is reasonable since for a D/M/1 queue the time-contraction algorithm simulates a change in arrival rate in the exact way dictated by the structure of the queue.

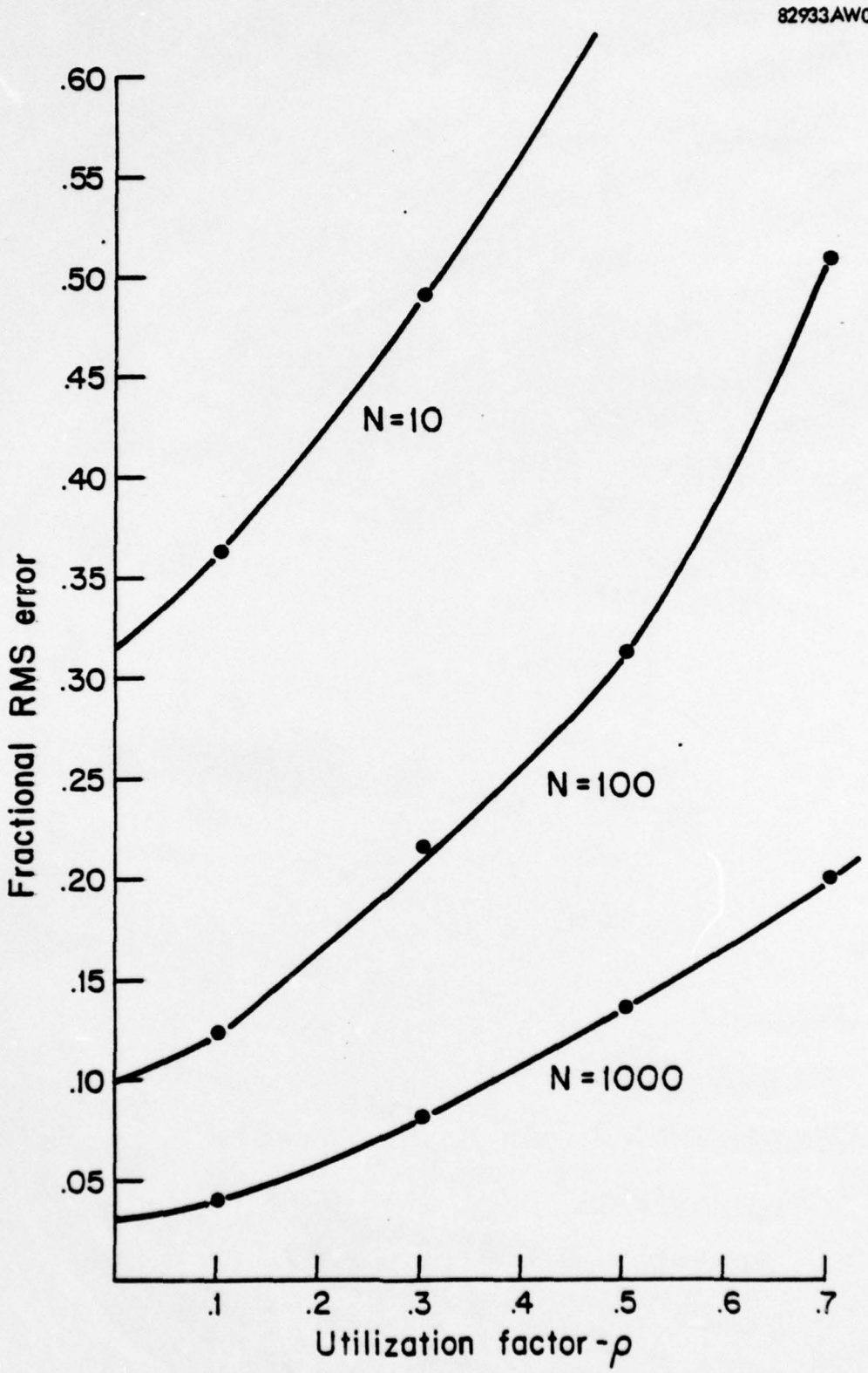
Hence, we have three estimation algorithms that appear from simulation and analytical results to be asymptocially unbiased, consistent, and efficient in the case of an M/D/1 queue. In evaluating the most promising algorithm as far as robustness, computational complexity, and storage requirements, we can only choose between the customer-removal and time-contraction algorithms, since the customer-addition procedure as formulated is only applicable to queues where all customers have identical service requirements. The simulation results suggest the time-contraction algorithm as the best candidate since it appears to be at least as robust as the customer-removal procedure, while having considerably less computational and storage requirements.



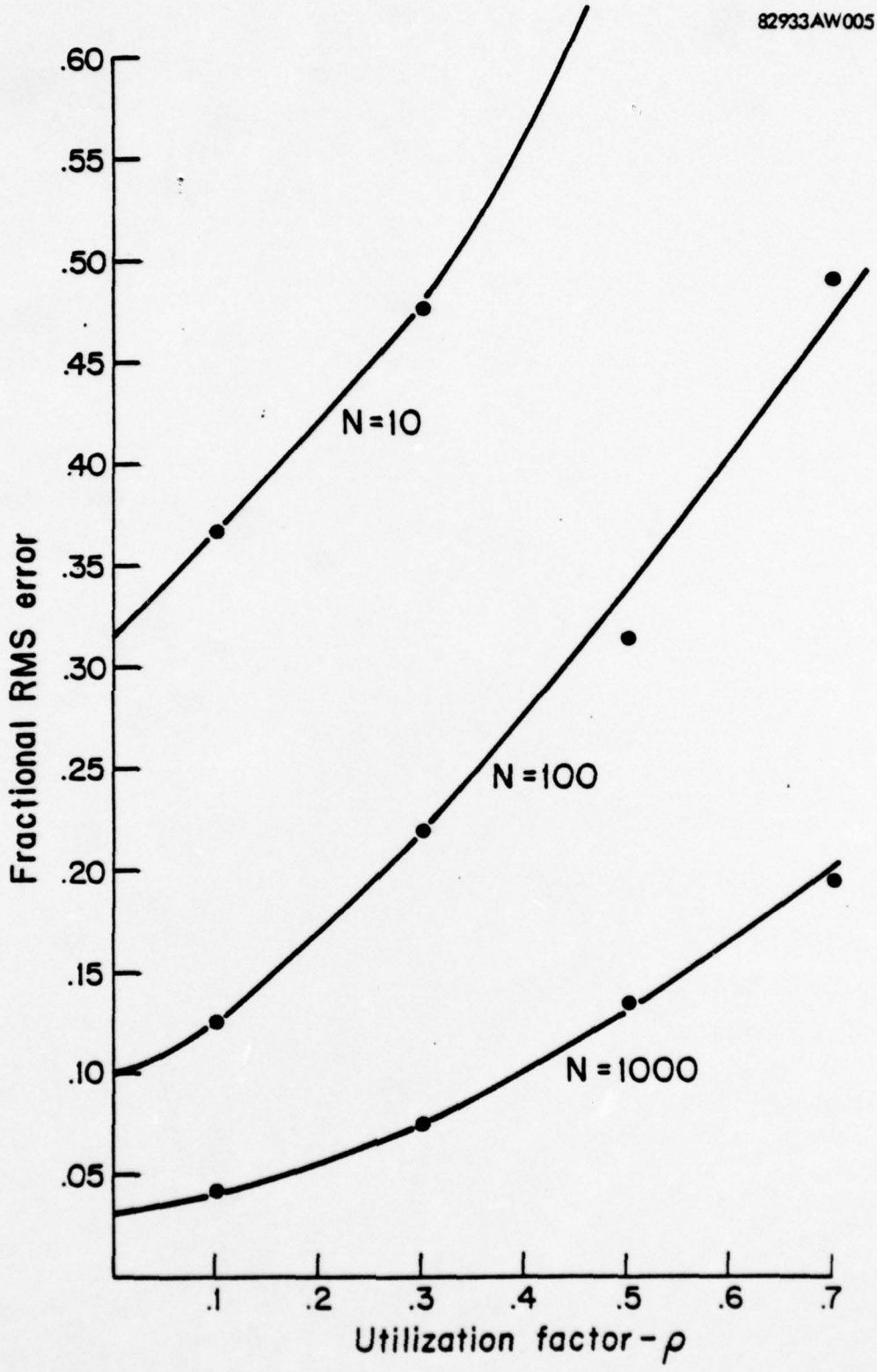
Graph (5.3) Time-Contraction Algorithm - Fractional RMS Errors for M|M|1.



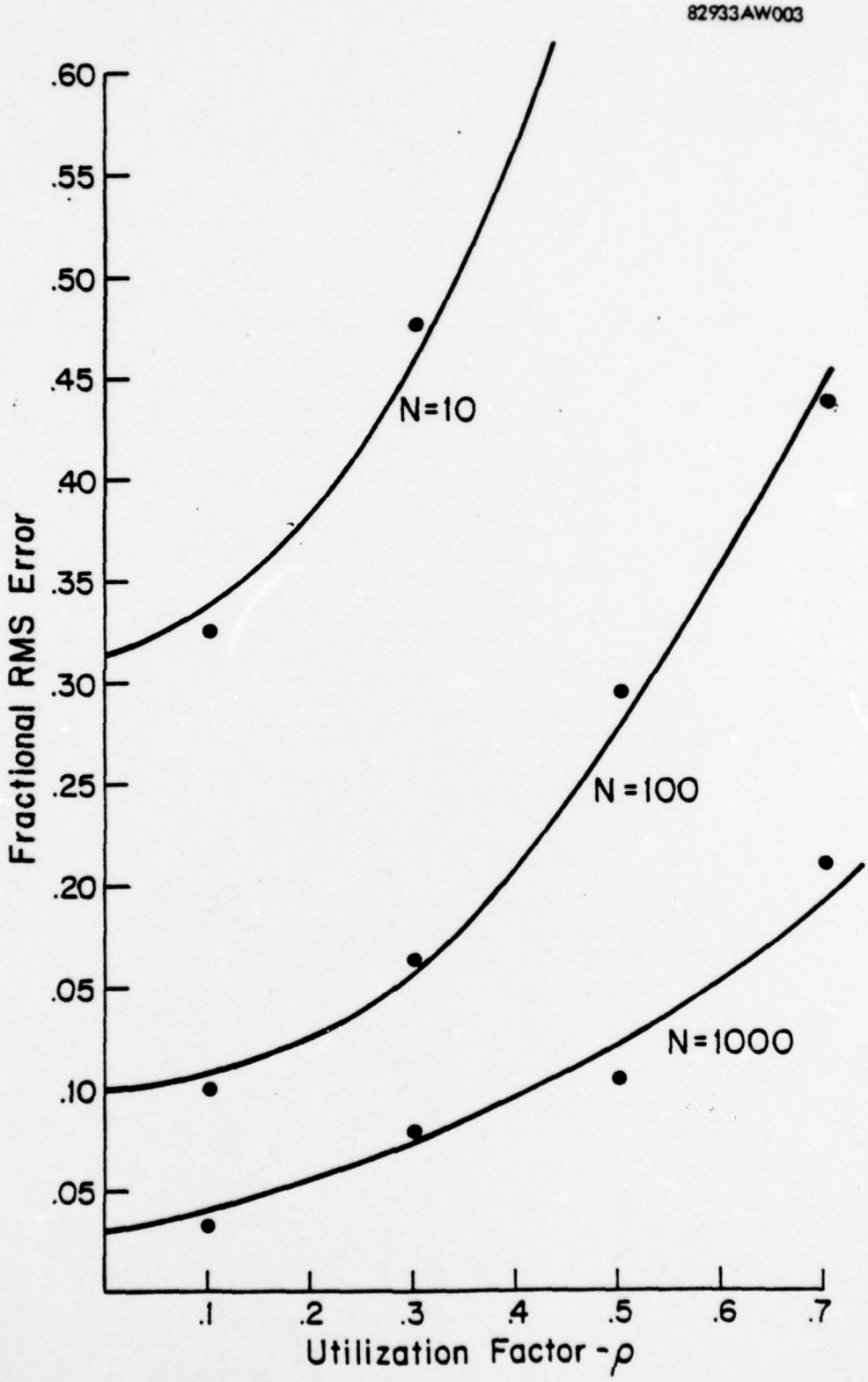
Graph (5.4) Customer-Removal Algorithm - Fractional RMS Errors for M|M|1.



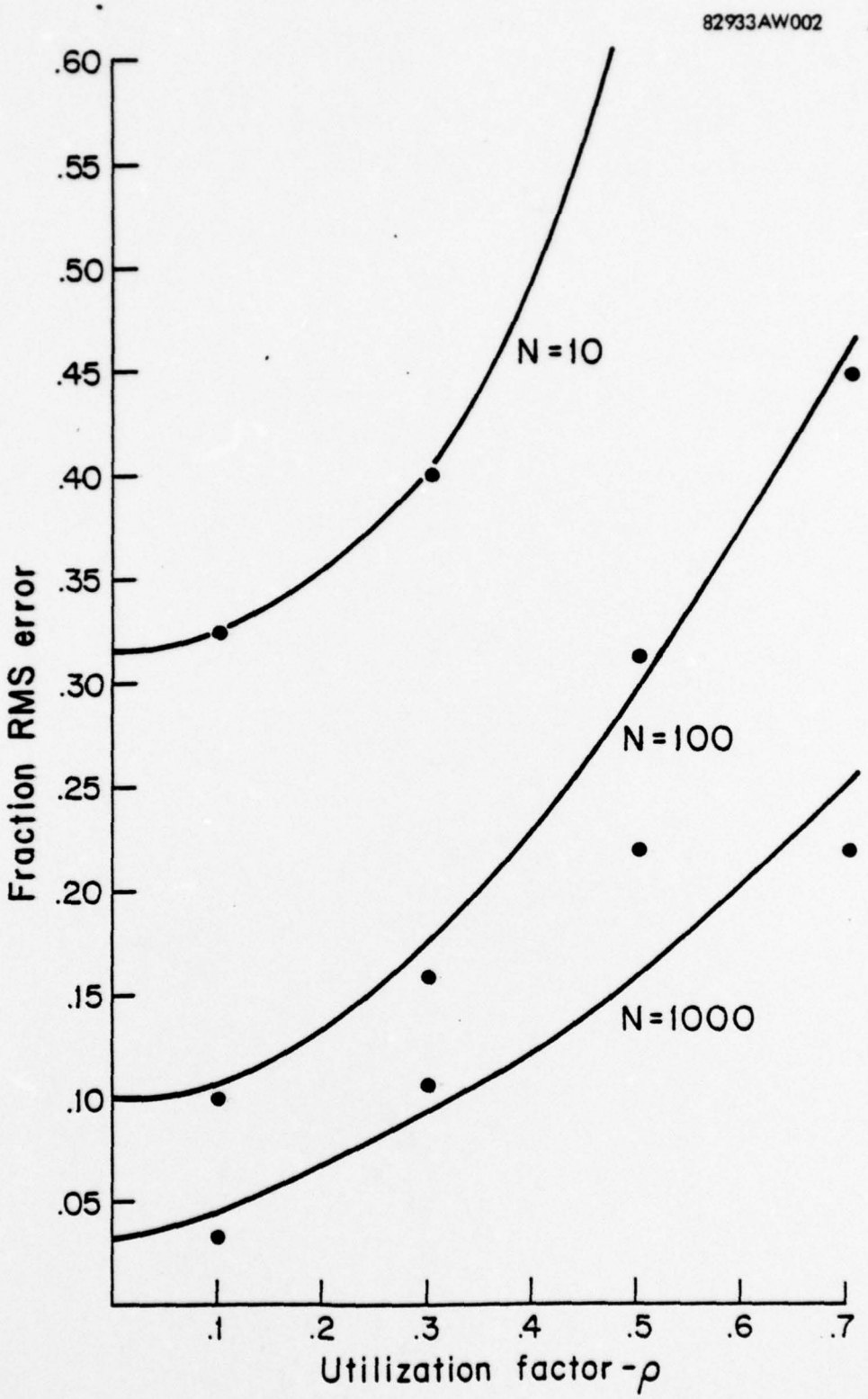
Graph(5.5) Time-Contaction Algorithm- Fractional RMS Error for U|M|1.



Graph(5.6) Customer-Removal Algorithm - Fractional RMS Error for U|M|1.



Graph(5.7) Time-Contraction Algorithm - Fractional RMS Error for D|M|1.



Graph (5.8) Customer-Removal Algorithm - Fractional RMS Error for D|M|1.

6. REFERENCES

[1] A. Segall, "The Modeling of Adaptive Routing in Data-Communication Networks," IEEE Trans. on Commun., V. Com-25, No. 1, January 1977 (pp. 85-95).

[2] R.G. Gallager, "A Minimum Delay Routing Algorithm Using Distributed Computation," IEEE Trans. on Commun., V. Com-25, No. 1, January 1977 (pp. 73-85).

[3] A. Segall, "Optimal Distributed Routing for Virtual Line-Switched Data Networks," IEEE Trans. on Commun., V. Com-27, No. 1, January 1979 (pp. 201-208).

[4] C. Agnew, "On the Optimality of Adaptive Routing Algorithms," in Conf. Rec. Nat. Telecommun. Conf. 1974 (pp. 1021-1025).

[5] C. Agnew, "On Quadratic Adaptive Routing Algorithms," Commun. Ass. Comput. Mach., V. 19, No. 1, 1976 (pp. 18-22).

[6] A. Ephremides, "Extension of an Adaptive Routing Algorithm to Mixed Media Networks," IEEE Trans. on Commun., V. Com-26, No. 8, August 1978 (pp. 1262-1266).

[7] L. Kleinrock, Queueing Theory V. 1, Wiley-Interscience, 1975.

[8] M. Bello, M.S. Thesis (The Estimation of Delay Gradients for Purposes of Routing in Data-Communication Networks), Report ESL-R-755.

imates for the mean and variance of the

busy periods, $\hat{D}'_{(k)}$, we generate a

iod records, processing each to form

We compute estimates of the bias

ith $\hat{D}'_{(k)}$ as follows:

, (5.2)

$$\text{k),i} - \left(\frac{1}{N_s} \sum_{i=1}^{N_s} \hat{D}'_{(k),i} \right)^2 . \quad (5.3)$$

ost interested in is the fractional

(5.4)